CEST 2017

15th International Conference on Environmental Science And Technology
31st August – 2nd September 2017, Rhodes, Greece

Stochastic simulation of periodic processes with arbitrary marginal distributions

Tsoukalas Ioannis, Efstratiadis Andreas and Makropoulos Christos

Department of Water Resources and Environmental Engineering
School of Civil Engineering
National Technical University of Athens, Greece
e-mail: itsoukal@mail.ntua.gr

*Submitted for CEST2017 young researchers award

Why stochastic hydrology?

Widely applied for (among others):

- Time series forecasting,
- Filling of missing records and
- Synthesis of long hydrologic time series that resemble the observed statistical characteristics.

Synthetic hydrology is of particular importance in water-related studies since it:

- Enables to account for the intrinsic uncertainty of hydrologic variables (e.g., precipitation and streamflow)
- Provides the means to uncertainty-proof the decision-making process of design and operation of a water-systems [*Matalas,* 1967].
- Allows the establishment of probability- and reliability-based methods and analysis.

Synthetic hydrology (contd.)

Usually credited to the pivotal research conducted by the Harvard water program [*Maass et al.,* 1962] and *Thomas and Fiering* [*1962*].

Characteristics of hydrological time series

- Long-range dependency (usually observed at annual or over-annual time scales).
- Intermittency (usually observed at fine time scales, e.g., daily).
- Non-Gaussianity (usually observed at all time scales!).
- Periodicity (usually observed at monthly time scale), i.e., Periodic fluctuations of the marginal statistics of the underlying process as well as a periodic correlation structure.

The standard hypothesis

- Generation of synthetic time series that preserve the essential statistical characteristics (marginal and joint) of the corresponding historical data [*Matalas and Wallis,* 1976; *Salas,* 1993].
- Yet, often **without** paying **attention** to the preservation of **their distribution**.

However, as emphatically remarked by *Klemeš and Borůvka* [1974]:

“*Simulation of a serially **correlated** series with a **given marginal distribution** is one of the **important prerequisites of synthetic hydrology** and of its applications to analysis of water resource system*”.

Current popular approaches

Transformation-based approaches

1) Typically require the use of complicated functions with many parameters. Which inevitably leads to the next questions:

   How many parameters are enough?

   How does the sample size effects their estimation?

   Are they robust enough?

2) Implications in cyclo-stationary simulation problems. Which rises questions such as:

   Should we use the same transformation function for all seasons? If yes, is it equally suitable?

   In what extent season-season correlation coefficients effected?

3) Introduction of bias in the resulting marginal statistics and stochastic structure of the process [*Salas et al.,* 1980 p. 73].

4) The resulting marginal distribution is difficult to identify *a priori* (it may not belong to a certain known family of distributions).

Use of non-Gaussian innovations for the white noise*

1) Except the Gaussian case, they are limited into **approximating** specific distribution families (usually, Pearson Type-3 or Log-Normal).

2) They are typically limited in models of order 1 – thus accounting only for lag-1 autocorrelation.

3) They are prone to generation of negative values, which are **not** appropriate for hydrologic processes.

4) They require the generation of innovation variables with higher skewness coefficients (due to central limit theorem). It is known that high values of skewness may cause a series of problems [e.g., *Todini*, 1980; *Koutsoyiannis*, 1999].

5) The synthesized time series resemble the historical statistics **but not** the **target marginal distributions**.

6) They can lead in bounded & unrealistic dependency forms.

* Throughout discussion and demonstration in a subsequent simulation study.

Stochastic Periodic AutoRegressive To Anything (SPARTA) model

Key idea of SPARTA:

Simulation of an auxiliary Periodic AutoRegressive (PAR) standard Gaussian process $\{Z_s\}$; $Z_s \sim N(0,1)$; where s refers to season; with such parameters (which define the stochastic structure) that after the mapping with the corresponding inverse cumulative distribution function (ICDF) results into a process $\{X_s\}$ with the desired (i.e., target) season-to-season correlation structure and marginal distributions. i.e.,

$$X_s = F_s^{-1}[\Phi(Z_s)]$$

Where $\Phi(\cdot)$ refers to the standard normal cumulative distribution function (CDF) and $F_s^{-1}(\cdot)$ denotes the ICDF of the target distribution of season s.

Origin:

Nataf's joint distribution model [*NDM; Nataf,* 1962], also known as NORmal To Anything procedure [NORTA; *Cario and Nelson,* 1997].

- Proposed for the generation of **correlated but serially independent** random vectors with given marginal distributions.

Rationale of NDM/NORTA:

- Employ an auxiliary multivariate Gaussian distribution with such parameters (i.e., correlation matrix) and subsequently map the generated data via the target ICDF. The resulting data will have the target marginal distributions and correlation structure.

Stochastic Periodic AutoRegressive To Anything (SPARTA) model

Main challenge:

- Identification of the parameters of the auxiliary process that result in the desired stochastic structure after the application of the ICDF.
- This arises from the fact that **Pearson correlation coefficient** (which is used within the parameter identification procedure of linear stochastic models, such as PAR) is **not invariant under monotonic transformations**; such as those imposed by the ICDF [*Embrechts et al.,* 1999 p. 7].
- Therefore, we have to identify the "equivalent" correlation coefficients that should be used within the parameter identification procedure of the auxiliary PAR model in order to attain the target correlation coefficients.

SPARTA in a nutshell

Let assume that we wish to describe a SPARTA process $\{X_s\}$ of order 1 with each season s characterized by distribution function $\mathrm{P}(X_s \leq x_s) = F_{X_s}$ and season-to- season correlation $\rho_{s,s-1}$. The generation equation of the auxiliary standard Gaussian PAR(1) model is given by:

$$Z_s = \hat{\rho}_{s,s-1} Z_{s-1} + \sqrt{1 - \hat{\rho}_{s,s-1}{}^{2}}\, W_s$$

Where, $\hat{\rho}(\cdot)$ stands for the "equivalent" correlation coefficient and W_s is an i.i.d. random variable from $N \sim (0, 1)$.

For notational purposes allow us to define the indices, $X_i \coloneqq X_s$ and $X_j \coloneqq X_{s-1}$.

The season-to-season **correlation structure** of the $\{Z_s\}$ process is **associated** with that of $\{X_s\}$ since $X_s = F_{X_s}^{-1}[\Phi(Z_s)]$, i.e.,

$$\rho_{i,j} = \mathrm{Corr}\left[X_i, X_j\right] = \mathrm{Corr}\left\{F_i^{-1}[\Phi(Z_i)], F_j^{-1}\left[\Phi\left(Z_j\right)\right]\right\}$$

By definition of Pearson's correlation coefficient:

$$\rho_{i,j} = \mathrm{Corr}\left[X_i, X_j\right] = \frac{E\left[X_i, X_j\right] - E[X_i]E\left[X_j\right]}{Var[X_i]Var\left[X_j\right]}$$

Known from the corresponding distributions F_{X_i} and F_{X_j}

Using the first Cross-product moment of X_i, X_j:

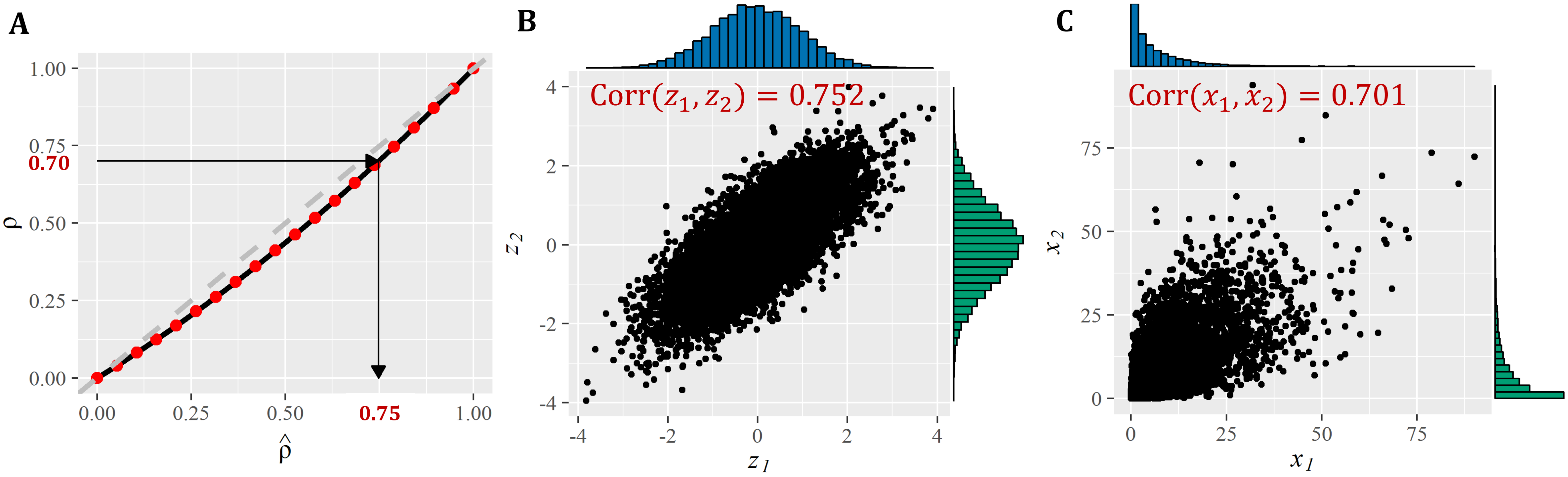
$$E\left[X_i, X_j\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{X_i}^{-1}[\Phi(z_i)] F_{X_j}^{-1}\left[\Phi\left(z_j\right)\right]\ \varphi\left(z_i, z_j, \hat{\rho}_{i,j}\right) \mathrm{d}z_i \mathrm{d}z_j$$

$$\rho_{i,j} = \mathrm{Corr}\left[X_i, X_j\right] = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{X_i}^{-1}[\Phi(z_i)] F_{X_j}^{-1}\left[\Phi\left(z_j\right)\right]\ \varphi\left(z_i, z_j, \hat{\rho}_{i,j}\right) \mathrm{d}z_i \mathrm{d}z_j - E[X_i]E\left[X_j\right]}{Var[X_i]Var\left[X_j\right]} \quad \Longrightarrow \quad \rho_{i,j} = f\left(\hat{\rho}_{i,j}\right)$$

Where, $\varphi(z_i, z_j, \hat{\rho}_{i,j})$ is the bivariate normal probability density function (PDF) with correlation $\hat{\rho}_{i,j}$.

Target Vs Equivalent correlation coefficient

Let assume that we wish to generate two correlated random variables X_1 and X_2 with same ($F_{X_1} \equiv F_{X_2}$) target marginal distributions (i.e., Gamma distribution with *scale* = 10 and *shape* = 0.7) and correlation $\rho_{x_1,X_2} = 0.7$.



For further details regarding the resolution of the double infinite integral see for example, *Tsoukalas et al.,* [2017].

SPARTA generation mechanism

The methodology can be summarized in five steps:

Step 1: Define (i.e., fit) a suitable marginal distribution to each season.

Step 2: Select an appropriate auxiliary periodic Gaussian model (e.g., PAR(1)).

Step 3: Approximate the equivalent correlation of pairs of interest (e.g., those related with the model parameters).

Step 4: Estimate the parameters of the auxiliary process $\{Z_s\}$ using the equivalent correlations coefficients.

Step 5: Simulate a realization of the auxiliary process $\{Z_s\}$ and map the generated data to the real domain in order to attain the process $\{X_s\}$, using the ICDFs identified in step 1. i.e., $X_s = F_{\mathrm{s}}^{-1}[\Phi(Z_s)]$.

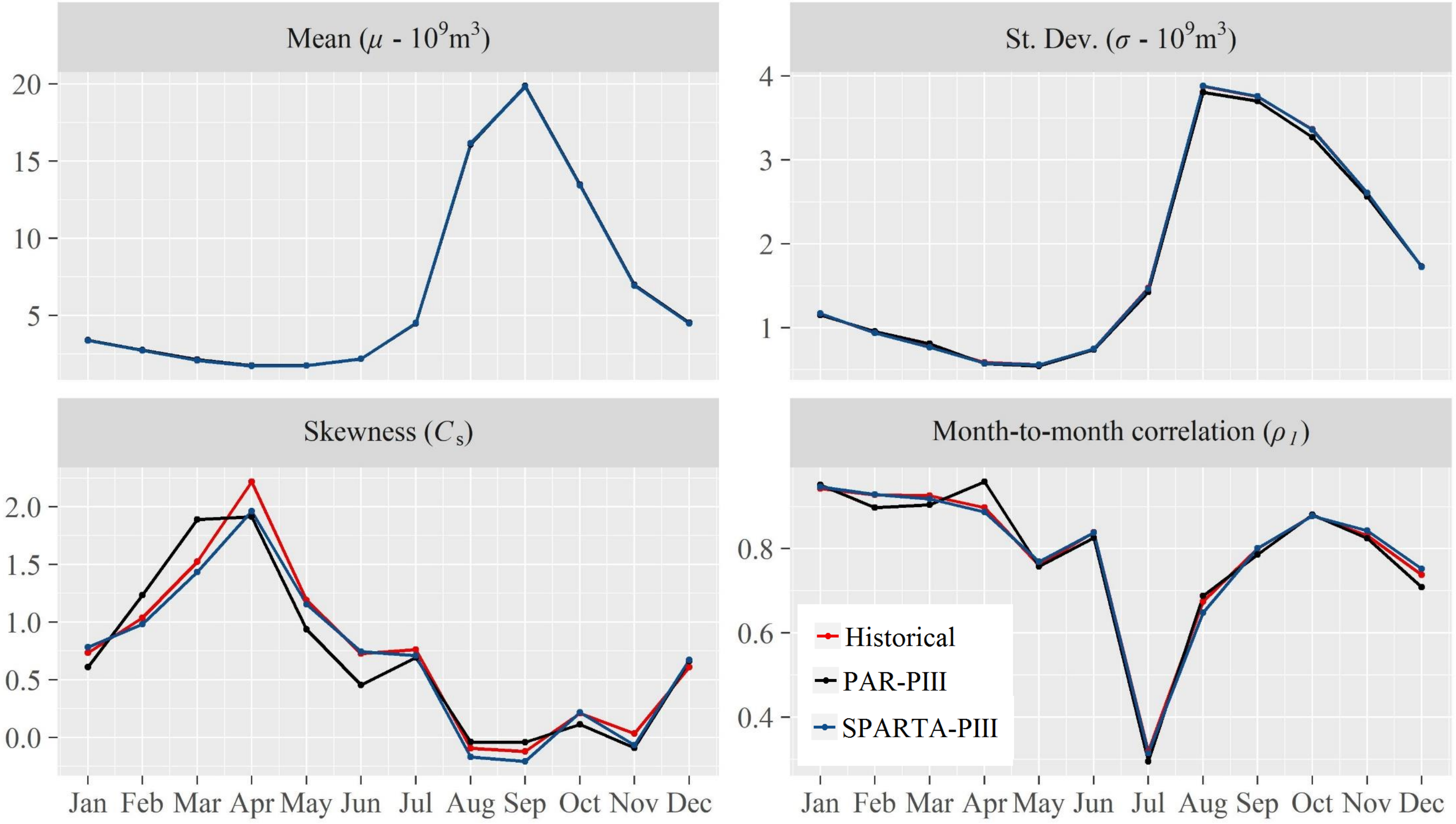
Application 1: Nile streamflow

Data[1]
Location: Nile River at Aswan dam.
Time step: Monthly.
Historical record extend: 1870 – 1945.
Characteristics: Skewed distributions, strong seasonality and high autocorrelations across all subsequent months.

Simulation study
Comparison with the typical PAR(1) model with Pearson type-III distribution for white noise (referred as PAR-PIII model. In order to conduct a fair and meaningful evaluation with SPARTA we also set this distribution as target one for all months (referred as SPARTA-PIII model).
Distribution: Pearson type-III fitted with the method of moments.
Synthetic time series length: 2 000 years.

[1] http://www.stats.uwo.ca/faculty/mcleod/epubs/mhsets/

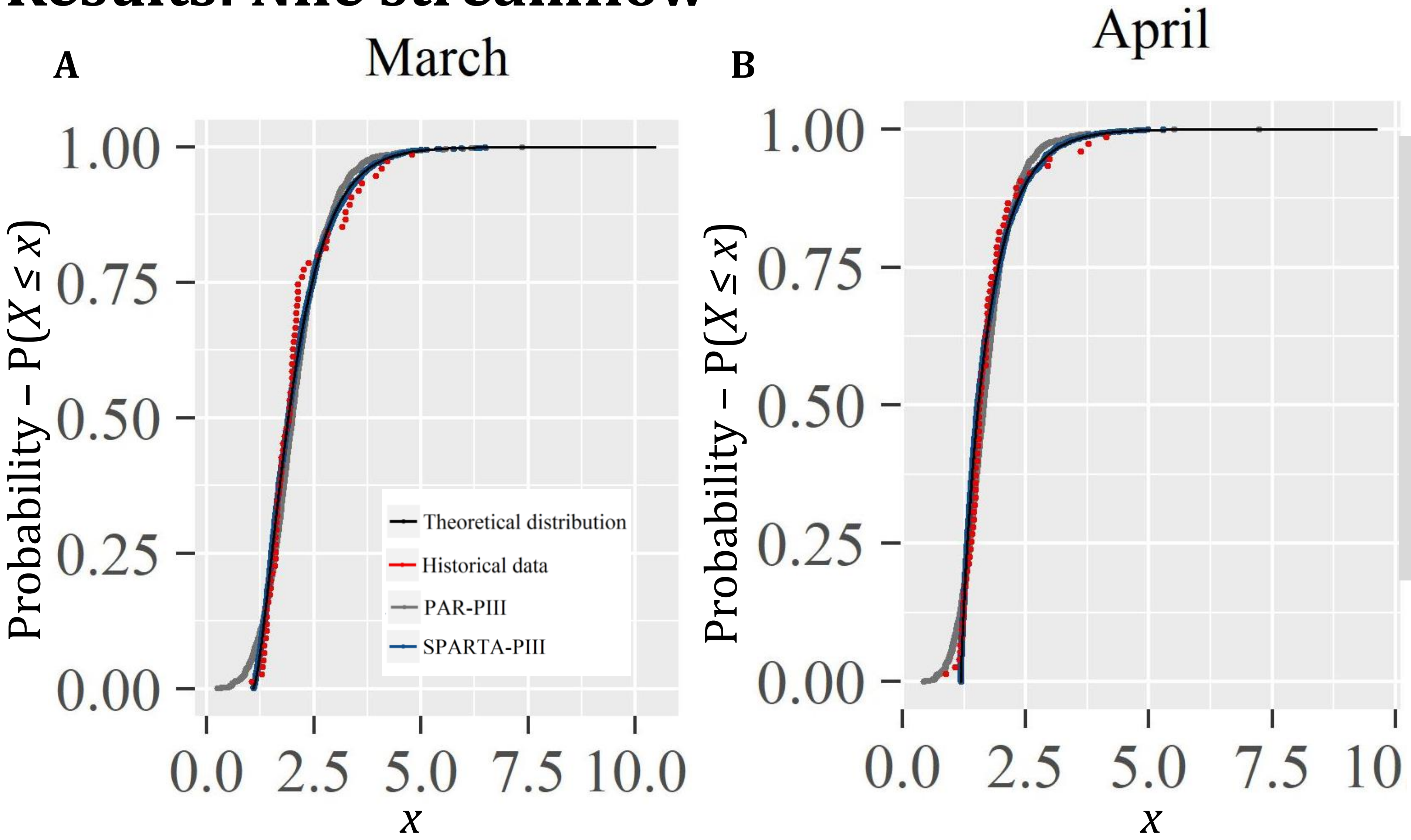
Results: Nile streamflow



Comparison between historical (**red line**) and simulated statistics with PAR-PIII (**black line**) and SPARTA-PIII (**blue line**) models.

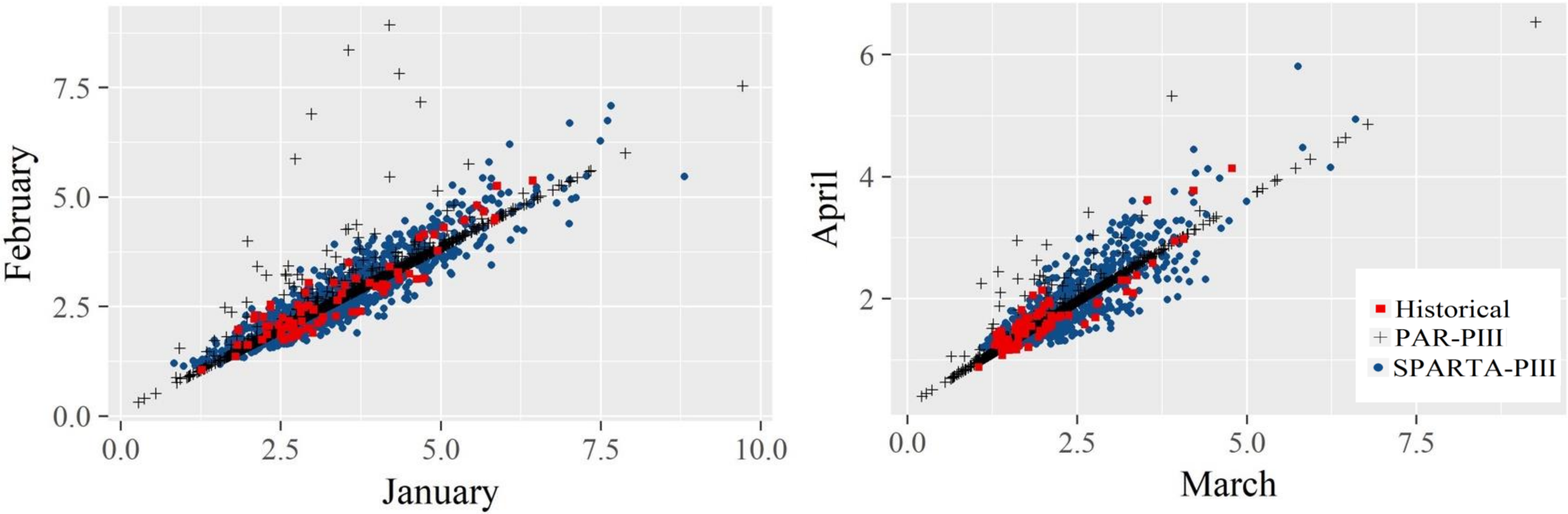
Tsoukalas et al., 2017. Stochastic simulation of periodic processes with arbitrary marginal distributions -

Results: Nile streamflow



Comparison of theoretical (**black line**), historical (**red line**) and simulated with PAR-PIII (**Grey line**) and SPARTA-PIII (**blue line**) models **cumulative density function (CDF)** of **A**) March and **B**) April. (Weibull plotting position).

Results: Nile streamflow



Comparison in terms of scatter plot between the historical (red squares) and simulated dependency patterns established with PAR-PIII (**black crosses**) and SPARTA-PIII (**blue dots**) models.

Application 2: Hypothetical simulation study assuming different distribution for each season

Theoretical distributions and parameters of each season of the artificial time series as well as MLE estimation of simulated data.

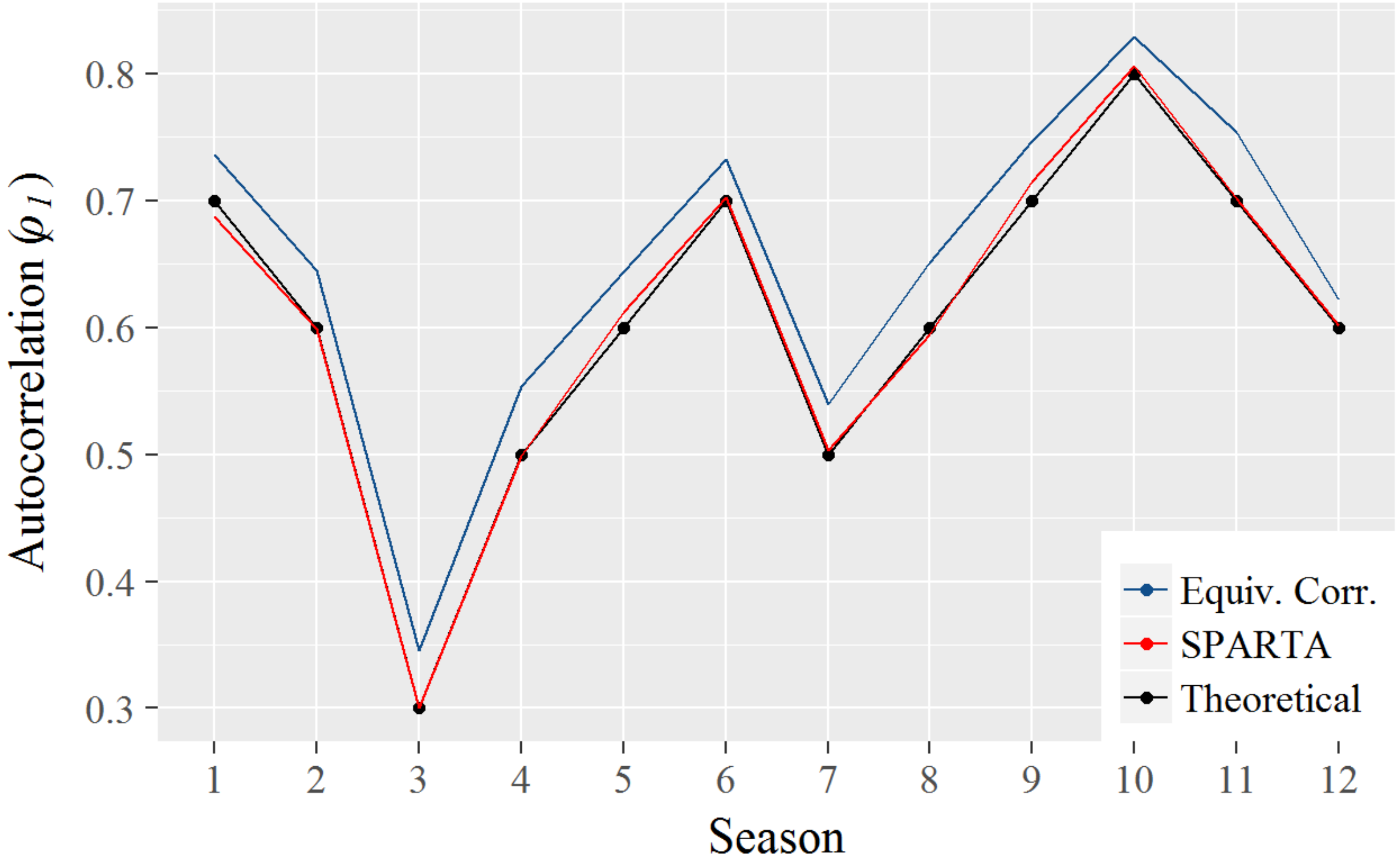
Season	1	2	3	4	5	6	7	8	9	10	11	12
Distribution/	PIII	Exp	Gam	Norm	LoNo	Wei	Beta	LoNo	Exp	PIII	Wei	Gam
Parameters	Theoretical Values											
a	1	1	1	2	0	1	1	0	0.55	1	2.5	2
b	2	-	2	1	0.5	2	5	0.7	-	1	5	1
c	2	-	-	-	-	-	-	-	-	5	-	-
	Simulated Values											
a	1.01	0.97	1.02	1.97	0.001	1.04	1.98	0.002	0.52	1.01	2.48	2.02
b	1.97	-	2.01	0.99	0.50	2.02	4.92	0.71	-	0.97	5.01	1.02
c	2.05	-	-	-	-	-	-	-	-	5.03	-	-

*Distribution abbreviations: PIII: Pearson III (a = shape, b = rate, c = location), Exp: Exponential (a = rate), Gam: Gamma (a = shape, b = rate), Norm: Normal (a = mean, b = st. dev.), LoNo: Log-Normal (a = log mean, b= log st. dev.), Wei: Weibull (a = shape, b = scale); Beta: Beta (a = shape, b = shape).

Furthermore, we assumed that the target season-to-season correlation is equal to:

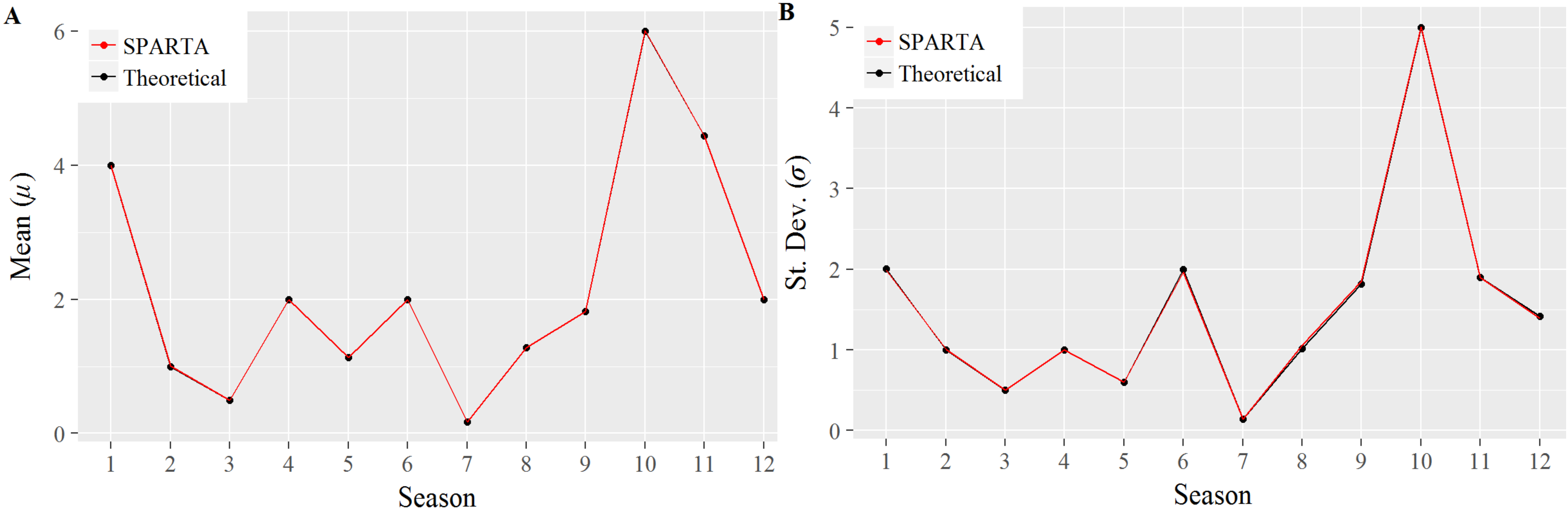
$$\boldsymbol{\rho} = \left[\rho_{12,1}, \rho_{1,2}, \dots, \rho_{t,t-1} \dots, \rho_{11,12}\right] = [0.7, 0.6, 0.3, 0.5, 0.6, 0.7, 0.5, 0.6, 0.7, 0.8, 0.7, 0.6].$$

Results: Hypothetical simulation study



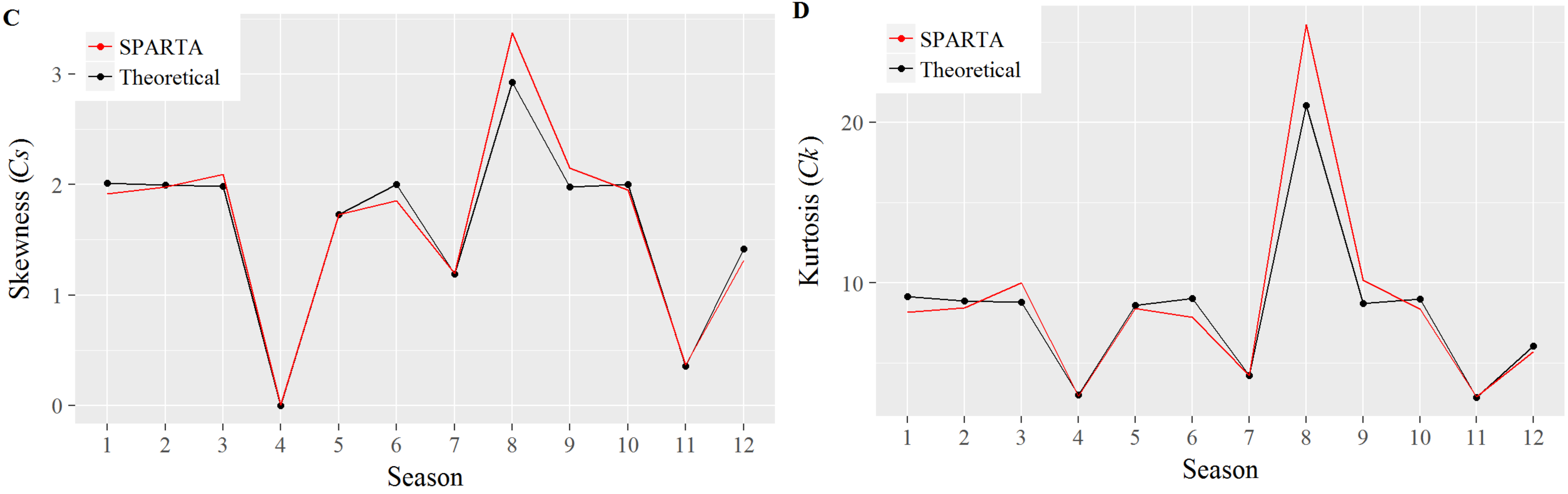
Comparison between theoretical (**black line**) and simulated (**red line**) lag-1 **season-to-season correlation** (ρ_1). The **blue line** illustrated the estimated equivalent correlation coefficients.

Results: Hypothetical simulation study



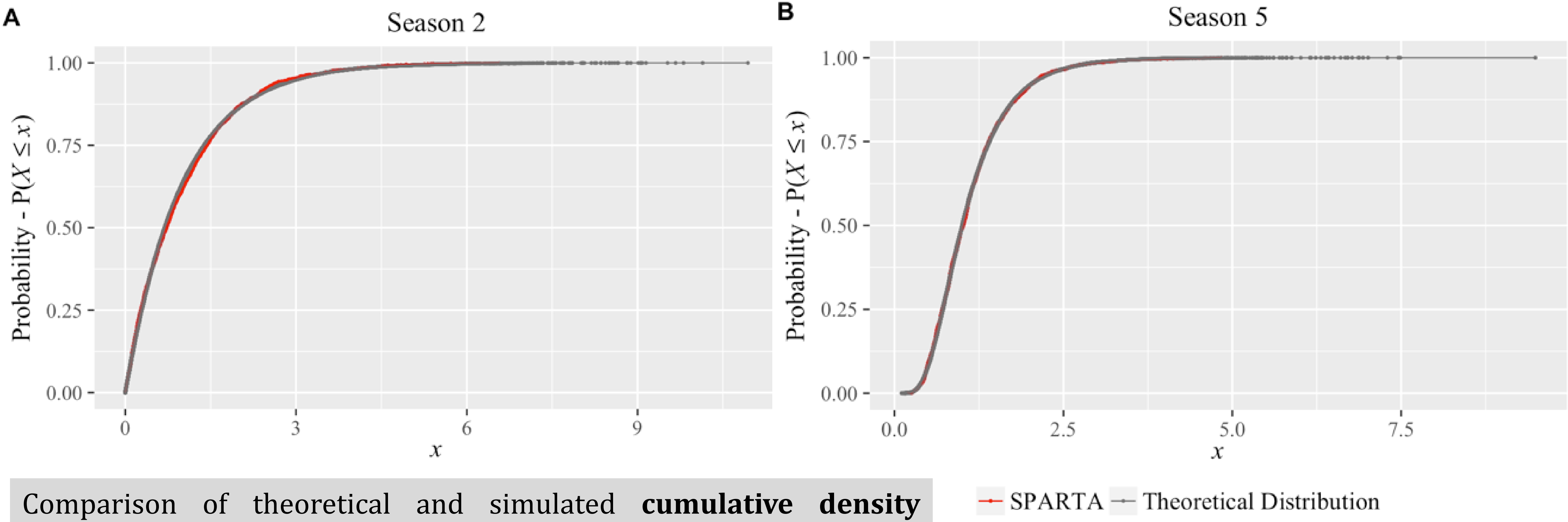
Comparison of theoretical and simulated values of seasonal **A) mean** (μ) and **B) standard deviation** (σ).

Results: Hypothetical simulation study



Comparison of theoretical and simulated values of seasonal **C**) **skewness** (C_s) and **D**) **kurtosis** (C_k) coefficients.

Results: Hypothetical simulation study



Comparison of theoretical and simulated **cumulative density function (CDF)** of **A**) season 2 and **B**) season 5 (Weibull plotting position).

Tsoukalas et al., 2017. Stochastic simulation of periodic processes with arbitrary marginal distributions

Conclusions

Advantages:

- Simulation of processes exhibiting:
  - Periodicity (i.e., cyclo-statrionary).
  - Any marginal distribution.
- Parsimonious structure and straightforward application.
- Tackle a series of problems relevant with the current approaches, such as,
  - Avoid the generation of negative values.
  - Generation of realistic dependency patterns.
  - Straightforward preservation of skewness coefficient.
- Incorporation of recent advances in statistical science into stochastic modelling.
  - e.g., robust distribution fitting methods (e.g., L-moments or maximum likelihood).

Ongoing and future work [Submitted in Water Resources Research, *Tsoukalas et al.,* 2017]:

- Extend SPARTA for multivariate time series simulation.
- Incorporate SPARTA within a disaggregation framework in order to account for long-range dependency and Hurst-Kolmogorov behavior as well as multi-scale consistency.
- Time series forecasting (e.g., streamflow, precipitation).

References

- Cario, M.C., and Nelson, B.L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Ind. Eng. 1–19.
- Embrechts, P., McNeil, A. J., & Straumann, D., (1999). Correlation and Dependence in Risk Management: Properties and Pitfalls, in *Risk Management*, edited by M. A. H. Dempster, pp. 176–223, Cambridge University Press, Cambridge.
- Fiering, B., and Jackson, B., (1971). Synthetic Streamflows, Water Resources Monograph, American Geophysical Union, Washington, D. C.
- Klemeš, V., and Borůvka, L., (1974). Simulation of Gamma-Distributed First-Order Markov Chain, *Water Resour. Res.*, *10*(1), 87–91, doi:10.1029/WR010i001p00087.
- Koutsoyiannis, D. (1999). Optimal decomposition of covariance matrices for multivariate stochastic models in hydrology, *Water Resources Research*, *35*(4), 1219–1229.
- Maass, A., Hufschmidt, M. M., Dorfman, R., Thomas, H. A., Marglin, S. A., Fair, G. M., Bower, B. T., Reedy, W. W., Manzer, D. F., & Barnett, M. P., (1962). *Design of water-resource systems*, Cambridge: Harvard University Press.
- Matalas, N. C. (1967). Mathematical assessment of synthetic hydrology, *Water Resources Research*, *3*(4), 937–945, doi:10.1029/WR003i004p00937.
- Matalas, N. C., and Wallis, J. R., (1976). *Generation of synthetic flow sequences, Systems Approach to Water Management*, edited by A. K. Biswas, McGraw-Hill, New York, New York.
- Nataf, A. (1962). Statistique mathematique-determination des distributions de probabilites dont les marges sont donnees, *C. R. Acad. Sci. Paris*, *255*(1), 42–43.
- Salas, J. D. (1993). Analysis and modeling of hydrologic time series, in *Handbook of hydrology*, edited by D. R. Maidment, p. Ch. 19.1-19.72, Mc-Graw-Hill, Inc.
- Salas, J. D., Delleur, J. W. , Yevjevich, V. , & Lane, W. L. (1980). *Applied modeling of hydrologic time series*, 2nd Print., Water Resources Publication, Littleton, Colorado.
- Todini, E. (1980). The preservation of skewness in linear disaggregation schemes, *Journal of Hydrology*, *47*(3–4), 199–214, doi:10.1016/0022-1694(80)90093-1.
- Tsoukalas, I., Efstratiadis, A., & Makropoulos, C. (2017). Stochastic periodic autoregressive to anything (SPARTA): Modeling and simulation of cyclostationary processes with arbitrary marginal distributions. Water Resources Research, 53. https://doi.org/10.1002/2017WR021394